

Learning the Irreducible Representations of Commutative Lie Groups

Taco Cohen

Machine Learning Group, University of Amsterdam

T.S.COHEN@UVA.NL

Max Welling

Machine Learning Group, University of Amsterdam

M.WELLING@UVA.NL

Abstract

We present a new probabilistic model of commutative Lie groups, that produces invariant-equivariant and disentangled representations of data. We borrow a fundamental principle from physics, used to define the elementary particles of a physical system, and use it to give a mathematically precise definition of the popular but heretofore rather vague notion of a “disentangled representation”. Our model is based on a new-found Bayesian conjugacy relation that enables us to perform fully tractable probabilistic inference over so-called Toroidal Lie groups – a class that includes practically relevant groups such as rotations and translations of images. We train the model on pairs of transformed image patches, and show that it produces a completely invariant representation which is highly effective for classification.

1. Introduction

Recently, the field of deep learning has produced some remarkable breakthroughs. The hallmark of the deep learning approach is to learn multiple layers of *representation* of data, and much work has gone into the development of representation learning modules such as RBMs and their generalizations (Welling et al., 2005), and autoencoders (Vincent et al., 2008). However, at this point it is not quite clear what makes a good representation. In this paper, we take a fresh look at the basic principles behind unsupervised representation learning from the perspective of Lie group theory¹.

¹We will at times assume a passing familiarity with Lie groups, but the main ideas of this paper should be accessible to a broad audience.

Various desiderata for learned representations have been expressed, including meaningful (Bengio & Lecun, 2014), invariant (Goodfellow et al., 2009), abstract and disentangled (Bengio et al., 2013) representations, but so far most of these notions have not been defined in a mathematically precise way. We focus on the notions of invariance and disentangling, leaving the search for meaning for future work.

What do we mean, intuitively, when we speak of invariance and disentangling? A disentangled representation is one that explicitly represents the distinct factors of variation in the data. For example, visual data can be thought of as a composition of object identity, position and pose, lighting conditions, etc. Once disentangling is achieved, invariance follows easily: to build a representation that is invariant to a transformation (e.g. a change in object position) that is considered a nuisance for a particular task (e.g. object classification), one can simply ignore the units in the representation that encode position.

To get a mathematical handle on the heretofore vague concept of disentangling, we borrow a fundamental principle from physics, which we refer to as Weyl’s principle, following Kanatani (1990). In physics, this idea is used to tease apart (i.e. disentangle) the *elementary particles* of a physical system from mere measurement values that have no inherent physical significance. We apply this principle to the area of vision, for after all, pixels are nothing but physical measurements.

Weyl’s principle presupposes a symmetry group that acts on the data. By this we mean a set of transformations that does not change the “essence” of the measured phenomenon, although it may change the “superficial appearance”, i.e. the measurement values. As a concrete example that we will use throughout this paper, consider the group known as $SO(2)$, acting on images by 2D rotation about the origin. A transformation from this group (a rotation) may change the value of every pixel in the image, but leaves invariant the identity of the imaged object. Weyl’s principle states that the elementary components of this system (the world as measured by a camera) are given by the *irre-*

ducible representations of the symmetry group – a concept that will be explained in this paper.

We demonstrate this theoretical principle using the so-called toroidal subgroups of the special orthogonal group as symmetry groups. This class includes several groups of practical interest, such as image rotations and translations. We introduce a probabilistic model, which we call Toroidal Subgroup Analysis (TSA), and show how it can be learned from pairs of images related by arbitrary and unobserved transformations in the group. Using a novel conjugate prior, the model integrates probability theory and Lie group theory in a very elegant way. All the relevant probabilistic quantities such as normalization constants, moments, KL-divergences, the posterior density over the transformation group, the marginal density in data space, and their gradients can be obtained in closed form. This complete tractability is a rather exceptional property among probabilistic graphical models with non-linear interactions.

The parameters of the posterior distribution over the transformation group provide a disentangled representation of the data that can be split into invariant and equivariant components. The structure of the learned group dictates a particular *pooling scheme* for computing invariants, reproducing a key feature of convolutional networks (LeCun & Bottou, 1998). When trained on shifts, the model learns to perform a Discrete Fourier Transform (DFT), thus providing a probabilistic interpretation of Fourier analysis and opening up the possibility of learning appropriate probabilistic generalizations of the DFT for a particular signal.

1.1. Related work

The first to propose an algorithm for learning Lie groups from data were Rao & Ruderman (1999). This work was later extended by Miao and Rao (2007). The first model deals only with one-parameter Lie groups, while the later work incorporates multiple transformation types. These works left open the problem of efficiently inferring transformation parameters for non-infinitesimal transformations. This problem was solved by Sohl-Dickstein et al. (2010) using an elegant adaptive smoothing technique, making it possible to learn from large transformations.

Other, non group theoretical approaches to learning transformations and invariant representations exist that do something similar to the irreducible reduction of a toroidal group, but this has not been recognized as a general principle for disentangling that is applicable to other groups, too. Gating models (Memisevic & Hinton, 2010) were found to perform a kind of eigenspace analysis (Memisevic, 2012), which is similar in spirit to the TSA model. However, gating models require far more parameters because they must tile each eigenspace with many pairs of filters whereas TSA

requires only two per eigenspace. Motivated by a number of statistical phenomena observed in natural images, Cadieu & Olshausen (2012) describe a model that decomposes a signal into invariant amplitudes and model transformations using phase variables.

None of the mentioned methods take into account the full uncertainty over transformation parameters, as does TSA. Because a unique transformation relating two images cannot in general be inferred, a multimodal posterior is required to give a complete description of the geometric situation. Furthermore, posterior inference in our model is very fast, whereas the MAP inference algorithm by Sohl-Dickstein et al. requires a more expensive iterative optimization.

2. Preliminaries

2.1. Equivalence, Invariance and Reducibility

In this section, we discuss three fundamental concepts on which the analysis in the rest of this paper is based: equivalence, invariance and reducibility.

Consider a function $\Phi : \mathbb{R}^D \rightarrow X$ that assigns to each possible data point $\mathbf{x} \in \mathbb{R}^D$ a class-label ($X = \{1, \dots, L\}$ being a discrete set of labels in this case) or some distributed representation (e.g. $X = \mathbb{R}^L$). Such a function induces an equivalence relation on the input space \mathbb{R}^D : we say that two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ are Φ -equivalent if they are mapped onto the same representation by Φ . Symbolically, $\mathbf{x} \equiv_{\Phi} \mathbf{y} \Leftrightarrow \Phi(\mathbf{x}) = \Phi(\mathbf{y})$.

Every equivalence relation on the input space fully determines a *symmetry group* acting on the space. This group, call it G , contains all transformations $\rho : \mathbb{R}^D \rightarrow \mathbb{R}^D$ that leave Φ invariant: $G = \{\rho \mid \Phi(\rho(\mathbf{x})) = \Phi(\mathbf{x})\}$. G describes the symmetries of Φ , or, stated differently, the label function/representation Φ is *invariant* to transformations in G acting on the input space. Hence, we can speak of G -equivalence: $\mathbf{x} \equiv_G \mathbf{y} \Leftrightarrow \exists \rho \in G : \rho(\mathbf{x}) = \mathbf{y}$. For example, if G is the group of 2D image rotations, two images are G -equivalent if they are rotations of each other.

Before we can introduce Weyl’s principle, we need one more concept: the *reduction* of a group representation. Let us assume that ρ acts linearly on our measurement space \mathbb{R}^D so that we can write $\rho(\mathbf{x}) = \mathbf{Q}\mathbf{x}$ for a matrix $\mathbf{Q} \in G$. In general, every coordinate y_i of $\mathbf{y} = \mathbf{Q}\mathbf{x}$ can depend on every coordinate x_j of \mathbf{x} . Since $\mathbf{x} \equiv_G \mathbf{y}$, it makes no sense to consider the coordinates x_i as separate quantities; we can only consider the vector \mathbf{x} as a unit because a symmetry transformation \mathbf{Q} tangles all coordinates.

However, we are free to change the basis of the measurement space. It may be possible to use a change of basis to expose an *invariant subspace*: a subspace $V \subset \mathbb{R}^D$ that

is mapped onto itself by every transformation in the group: $\forall \mathbf{Q} \in G : \mathbf{x} \in V \Rightarrow \mathbf{Q}\mathbf{x} \in V$. If such a subspace exists and its orthogonal complement $V^\perp \subset \mathbb{R}^D$ is also an invariant subspace, then it makes sense to consider the two parts of \mathbf{x} that lie in V and V^\perp be distinct quantities because they remain distinct under symmetry transformations.

Let \mathbf{W} be a change of basis matrix that exposes the invariant subspaces, that is,

$$\mathbf{Q} = \mathbf{W} \begin{bmatrix} \mathbf{Q}_1 & \\ & \mathbf{Q}_2 \end{bmatrix} \mathbf{W}^{-1}, \quad (1)$$

for all $\mathbf{Q} \in G$. The matrices \mathbf{Q}_1 and \mathbf{Q}_2 are functions of \mathbf{Q} , and each forms a representation of the *same* abstract group as represented by \mathbf{Q} . The group representations \mathbf{Q}_1 and \mathbf{Q}_2 describe how the individual parts $\mathbf{x}_1 \in V$ and $\mathbf{x}_2 \in V^\perp$ are transformed by the elements of the group. Following common practice, we refer to both the group representations \mathbf{Q}_1 , \mathbf{Q}_2 and the subspaces V and V^\perp corresponding to these group representations as “representations”.

The process of reduction can be applied recursively to \mathbf{Q}_1 and \mathbf{Q}_2 . If at some point there is no more (non-trivial) invariant subspace, the representation is called *irreducible*. Weyl’s principle states that *the elementary components of a system are the irreducible representations of the symmetry group of the system*. Properly understood, it is not a physics principle at all, but generally applicable to any situation where there is a well-defined notion of equivalence. It is completely abstract and therefore agnostic about the type of data (images, optical flows, sound, etc.), making it eminently useful for representation learning.

In the rest of this paper, we will demonstrate Weyl’s principle in the simple case of a compact commutative subgroup of the special orthogonal group (these terms are explained below). We want to stress though, that there is no reason the basic ideas cannot be applied to non-commutative groups acting on non-linear latent representation spaces.

2.2. Maximal Tori in the Orthogonal Group

In order to facilitate analysis, we will from here on consider only compact commutative subgroups of the special orthogonal group $SO(D)$. For reasons that will become clear shortly, such groups are called toroidal subgroups of $SO(D)$. Intuitively, the toroidal subgroups of general compact Lie groups can be thought of as the “commutative part” of these groups. This fact, combined with their analytic tractability (evidenced by the results in this paper) makes them suitable as the starting point of a theory of probabilistic Lie-group learning.

Imposing the constraint of orthogonality will make the computation of matrix inverses very cheap, because for orthogonal \mathbf{Q} , $\mathbf{Q}^{-1} = \mathbf{Q}^T$. Orthogonal matrices also avoid

numerical problems, because their condition number is always equal to 1. Another important property of orthogonal transformations is that they leave the Euclidean metric invariant: $\|\mathbf{Q}\mathbf{x}\| = \|\mathbf{x}\|$. Therefore, orthogonal matrices cannot express transformations such as contrast scaling, but they can still model the interesting structural changes in images (Bethge et al., 2007). For example, since 2D image rotation and (cyclic) translation are linear and do not change the total energy (norm) of the image, they can be represented by orthogonal matrices acting on vectorized images.

As is well known, commuting matrices can be simultaneously diagonalized, so one could represent a toroidal group in terms of a basis of eigenvectors shared by every element in the group, and one diagonal matrix of eigenvalues for each element of the group, as was done in (Sohl-Dickstein et al., 2010) for 1-parameter Lie groups. However, orthogonal matrices do not generally have a complete set of real eigenvectors. One could use a complex basis instead, but this introduces redundancies because the eigenvalues and eigenvectors of an orthogonal matrix come in complex conjugate pairs. For machine learning applications, this is clearly an undesirable feature, so we opt for a joint block-diagonalization of the elements \mathbf{Q} of the toroidal group: $\mathbf{Q}(\varphi) = \mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T$, where \mathbf{W} is orthogonal and $\mathbf{R}(\varphi)$ is a block-diagonal rotation matrix²:

$$\mathbf{R}(\varphi) = \begin{bmatrix} \mathbf{R}(\varphi_1) & & \\ & \ddots & \\ & & \mathbf{R}(\varphi_J) \end{bmatrix}. \quad (2)$$

The diagonal of $\mathbf{R}(\varphi)$ contains 2×2 rotation matrices

$$\mathbf{R}(\varphi_j) = \begin{bmatrix} \cos(\varphi_j) & -\sin(\varphi_j) \\ \sin(\varphi_j) & \cos(\varphi_j) \end{bmatrix} \quad (3)$$

In this parameterization, the real, orthogonal basis \mathbf{W} identifies the group representation, while the vector of rotation angles φ identifies a particular element $\mathbf{Q}(\varphi)$ of the group. It is now clear why such groups are called “toroidal”: the parameter space φ is periodic in each element φ_j and hence is a topological torus. For a J -parameter toroidal group, all the φ_j can be chosen freely. Such a group is known as a *maximal* torus in $SO(D)$, for which we write $\mathbb{T}^J = \{\varphi \mid \varphi_j \in [0, 2\pi], j = 1, \dots, J\}$.

To gain insight into the structure of toroidal groups with fewer parameters, we rewrite eq. 2 using the matrix exponential:

$$\mathbf{R}(\varphi) = \exp \left(\sum_{j=1}^J \varphi_j \mathbf{A}_j \right) \quad (4)$$

The the anti-symmetric matrices $\mathbf{A}_j = \frac{d}{d\varphi_j} \mathbf{R}(\varphi)|_0$ are

²For ease of exposition, we assume an even dimensional space $D = 2J$, but the equations are easily generalized.

known as the Lie algebra generators, and the φ_j are Lie-algebra coordinates.

The Lie algebra is a structure that largely determines the structure of the corresponding Lie group, while having the important advantage of forming a linear space. That is, all linear combinations of the generators belong to the Lie algebra, and each element of the Lie algebra corresponds to a transformation in the Lie group, which itself is a non-linear manifold. Furthermore, every subgroup of the Lie group corresponds to a subalgebra (not defined here) of the Lie algebra. All toroidal groups are the subgroup of some maximal torus, so we can learn a general toroidal group by first learning a maximal torus and then learning a subalgebra of its Lie algebra. Due to commutativity, the structure of the Lie algebra of toroidal groups is such that any subspace of the Lie algebra is in fact a subalgebra. The relevance of this observation to our machine learning problem is that *to learn a toroidal group with I parameters ($I < J$), we can simply learn a maximal toroidal group and then learn an I -dimensional linear subspace in the space of φ .*

In this work, we are interested in compact subgroups only³, which is to say that the parameter space should be closed and bounded. To see that not all subgroups of a maximal torus are compact, consider a 4D space and a maximal torus with 2 generators \mathbf{A}_1 and \mathbf{A}_2 . Let us define a subalgebra with one generator $\mathbf{A} = \omega_1 \mathbf{A}_1 + \omega_2 \mathbf{A}_2$, for real numbers ω_1 and ω_2 . The group elements generated by this algebra through the exponential map take the form

$$\mathbf{R}(s) = \exp(s\mathbf{A}) = \begin{bmatrix} \mathbf{R}(\omega_1 s) & \\ & \mathbf{R}(\omega_2 s) \end{bmatrix}. \quad (5)$$

Each $\mathbf{R}(\omega_j s)$ is periodic with period $2\pi/\omega_j$, but their combination $\mathbf{R}(s)$ need not be. When ω_1 and ω_2 are not commensurate, all values of $s \in \mathbb{R}$ will produce different $\mathbf{R}(s)$. To obtain a compact one-parameter group with parameter space $s \in [0, 2\pi]$, we restrict the frequencies ω_j to be integers, so that $\mathbf{R}(s) = \mathbf{R}(s + 2\pi)$ (see figure 1). To get a feel for what this means in the case of 2D image rotation, see figure 3: the radial frequency of each filter pair corresponds to the value ω_j for that pair.

It is easy to see that each block of $\mathbf{R}(s)$ forms an irreducible representation of the entire group that acts on the data. From the point of view expounded in section 2.1, we should view the vector \mathbf{x} as a tangle of *elementary components* $\mathbf{u}_j = \mathbf{W}_j^T \mathbf{x}$, where $\mathbf{W}_j = (\mathbf{W}_{(:, 2j-1)}, \mathbf{W}_{(:, 2j)})$ denotes the $D \times 2$ submatrix of \mathbf{W} corresponding to the j -th block in $\mathbf{R}(s)$. Each one of the elementary parts \mathbf{u}_j is functionally independent of the others under symmetry

³The main reason for this restriction is that compact groups are simpler and better understood than non-compact groups. In practice, many non-compact groups can be compactified, so not much is lost.

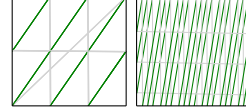


Figure 1. Parameter space $\varphi = (\omega_1 s, \omega_2 s)$ of two toroidal subgroups, for $\omega_1 = 2, \omega_2 = 3$ (left) and $\omega_1 = 1, \omega_2 = 2\pi$ (right). The point φ moves over the dark green line as s is changed. Wrapping around is indicated in light gray. In the incommensurable case (right), coupling does not add structure to the model, because all transformations (points in the plane) can still be constructed by an appropriate choice of s .

transformations.

The variable ω_j is known as the *weight* of the representation (Kanatani, 1990). When the representations are equivalent (i.e. they have the same weight), the parts are “of the same kind” and are transformed identically. Elementary components with different weights transform differently.

In the following section, we show how a maximal toroidal group and a 1-parameter subgroup can be learned from correspondence pairs, and how these can be used to generate invariant representations.

3. Toroidal Subgroup Analysis

We will start by modelling a maximal torus. A data pair (\mathbf{x}, \mathbf{y}) is related by a transformation $\mathbf{Q} = \mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T$ from the group representation:

$$\mathbf{y} = \mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T \mathbf{x} + \epsilon, \quad (6)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ represents isotropic Gaussian noise. In other symbols, $p(\mathbf{y}|\mathbf{x}, \varphi) = \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T \mathbf{x}, \sigma^2)$.

We will find it useful to introduce some notation for indexing invariant subspaces. As before, $\mathbf{W}_j = (\mathbf{W}_{(:, 2j-1)}, \mathbf{W}_{(:, 2j)})$. Let $\mathbf{u}_j = \mathbf{W}_j^T \mathbf{x}$ and $\mathbf{v}_j = \mathbf{W}_j^T \mathbf{y}$. If we want to access one of the coordinates of \mathbf{u} or \mathbf{v} , we write $u_{j1} = \mathbf{W}_{(:, 2j-1)}^T \mathbf{x}$ or $u_{j2} = \mathbf{W}_{(:, 2j)}^T \mathbf{x}$.

We assume the φ_j to be marginally independent and von-Mises distributed. The von-Mises distribution is an exponential family that assigns equal density to the endpoints of any length- 2π interval of the real line, making it a suitable choice for periodic variables such as φ_j . We will find it useful to move back and forth between the conventional and natural parameterizations of this distribution. The conventional parameterization of the von-Mises distribution $\mathcal{M}(\varphi|\mu, \kappa)$ uses a mean μ and precision κ :

$$p(\varphi_j) = \frac{1}{2\pi I_0(\kappa_j)} \exp(\kappa_j \cos(\varphi_j - \mu_j)). \quad (7)$$

The function I_0 that appears in the normalizing constant is known as the modified Bessel function of order 0.

Since the von-Mises distribution is an exponential family, we can write it in terms of natural parameters $\eta_j = (\eta_{j1}, \eta_{j2})^T$ as follows:

$$p(\varphi_j) = \frac{1}{2\pi I_0(\|\eta_j\|)} \exp(\eta_j^T T(\varphi_j)), \quad (8)$$

where $T(\varphi_j) = (\cos(\varphi_j), \sin(\varphi_j))^T$ are the sufficient statistics. The natural parameters can be computed from conventional parameters using,

$$\eta_j = \kappa_j [\cos(\mu_j), \sin(\mu_j)]^T \quad (9)$$

and vice versa,

$$\kappa_j = \|\eta_j\|, \quad \mu_j = \tan^{-1}(\eta_{j2}/\eta_{j1}) \quad (10)$$

Using the natural parameterization, it is easy to see that the prior is conjugate to the likelihood, so that the posterior $p(\varphi|\mathbf{x}, \mathbf{y})$ is again a product of von-Mises distributions. Such conjugacy relations are of great utility in Bayesian statistics, because they simplify sequential inference. To our knowledge, this conjugacy relation has not been described before. First observe that the likelihood term splits into a sum over the invariant subspaces indexed by j :

$$\begin{aligned} p(\varphi|\mathbf{x}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x}, \varphi) p(\varphi) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T \mathbf{x}\|^2\right) p(\varphi) \\ &\propto \exp\left(\sum_{j=1}^J \frac{\mathbf{v}_j^T \mathbf{R}(\varphi_j) \mathbf{u}_j}{\sigma^2} + \eta_j^T T(\varphi_j)\right) \end{aligned}$$

Both the bilinear forms $\mathbf{v}_j^T \mathbf{R}(\varphi_j) \mathbf{u}_j$ and the prior terms $\eta_j^T T(\varphi_j)$ are linear functions of $\cos(\varphi_j)$ and $\sin(\varphi_j)$, so that they can be combined into a single dot product:

$$p(\varphi|\mathbf{x}, \mathbf{y}) \propto \exp\left[\sum_{j=1}^J \hat{\eta}_j^T T(\varphi_j)\right], \quad (11)$$

which we recognize as a product of von-Mises in natural form.

The parameters $\hat{\eta}_j$ of the posterior are given by:

$$\begin{aligned} \hat{\eta}_j &= \eta_j + \frac{1}{\sigma^2} [u_{j1}v_{j1} + u_{j2}v_{j2}, u_{j1}v_{j2} - u_{j2}v_{j1}]^T \\ &= \eta_j + \frac{\|\mathbf{u}_j\| \|\mathbf{v}_j\|}{\sigma^2} [\cos(\theta_j), \sin(\theta_j)]^T, \end{aligned} \quad (12)$$

where θ_j is the angle between \mathbf{u}_j and \mathbf{v}_j . Geometrically, we can interpret the Bayesian updating procedure in eq. 12 as follows. The orientation of the natural parameter vector

η_j determines the mean of the von-Mises, while its magnitude determines the precision. To update this parameter with new information obtained from data $\mathbf{u}_j, \mathbf{v}_j$, one should add the vector $(\cos(\theta_j), \sin(\theta_j))^T$ to the prior, using a scaling factor that grows with the magnitude of \mathbf{u}_j and \mathbf{v}_j and declines with the square of the noise level σ . The longer \mathbf{u}_j and \mathbf{v}_j and the smaller the noise level, the greater the precision of the observation. This geometrically sensible result follows directly from the consistent application of the rules of probability.

Observe that when using a uniform prior (i.e. $\eta_j = \mathbf{0}$), the posterior mean $\hat{\mu}_j$ (computed from $\hat{\eta}_j$ by eq. 10) will be exactly equal to the angle θ_j between \mathbf{u}_j and \mathbf{v}_j . We will use this fact in section 3.1 when we derive the formula for the orbit distance in a toroidal group.

Previous approaches to Lie group learning only provide point estimates of the transformation parameters, which have to be obtained using an expensive iterative optimization procedure (Sohl-Dickstein et al., 2010). In contrast, TSA provides a full tractable posterior distribution which is obtained using a simple feed-forward computation. Compared to the work of Cadieu & Olshausen (2012), our model deals well with low-energy subspaces, by simply describing the uncertainty in the estimate instead of providing inaccurate estimates that have to be discarded.

3.1. Invariant Representation and Metric

One way of doing invariant classification is by using an invariant metric known as the *manifold distance*. This metric $d(\mathbf{x}, \mathbf{y})$ is defined as the minimum distance between the orbits $O_{\mathbf{x}} = \{\mathbf{Q}\mathbf{x} \mid \mathbf{Q} \in G\}$ and $O_{\mathbf{y}} = \{\mathbf{Q}\mathbf{y} \mid \mathbf{Q} \in G\}$. Observe that this is only a true metric that satisfies the coincidence axiom $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ if we take the condition $\mathbf{x} = \mathbf{y}$ to mean “equivalence up to symmetry transformations” or $\mathbf{x} \equiv_G \mathbf{y}$, as discussed in section 2.1.

In practice, it has proven difficult to compute this distance exactly, so approximations such as tangent distance have been invented (Simard et al., 2000). But for a maximal torus, we can easily compute the exact manifold distance:

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= \min_{\varphi} \|\mathbf{y} - \mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T \mathbf{x}\|^2 \\ &= \sum_j \min_{\varphi_j} \|\mathbf{v}_j - \mathbf{R}(\varphi_j) \mathbf{u}_j\|^2 \\ &= \sum_j \|\mathbf{v}_j - \mathbf{R}(\hat{\mu}_j) \mathbf{u}_j\|^2, \end{aligned} \quad (13)$$

where $\hat{\mu}_j$ is the mean of the posterior $p(\varphi_j|\mathbf{x}, \mathbf{y})$, obtained using a uniform prior ($\kappa_j = 0$). The last step of eq. 13 follows, because as we saw in the previous section, $\hat{\mu}_j$ is simply the angle between \mathbf{u}_j and \mathbf{v}_j when using a uniform prior. Therefore, $\mathbf{R}(\hat{\mu}_j)$ aligns \mathbf{u}_j and \mathbf{v}_j , thereby minimizing the distance.

Another approach to invariant classification is through an invariant representation. Although the model presented above aims to describe the transformation between observations \mathbf{x} and \mathbf{y} , an invariant-equivariant representation appears automatically in terms of the parameters of the posterior over the group. To see this, consider all the transformations in the learned toroidal group G that take an image \mathbf{x} to itself. This set is known as the stabilizer $\text{stab}_G(\mathbf{x})$ of \mathbf{x} . It is a subgroup of G and describes the symmetries of \mathbf{x} with respect to G . When a transformation $\mathbf{Q} \in G$ is applied to \mathbf{x} , the stabilizer subgroup is left invariant, for if $\mathbf{P} \in \text{stab}_G(\mathbf{x})$ then $\mathbf{P}(\mathbf{Q}\mathbf{x}) = \mathbf{Q}(\mathbf{P}\mathbf{x}) = \mathbf{Q}\mathbf{x}$ and hence $\mathbf{P} \in \text{stab}_G(\mathbf{Q}\mathbf{x})$.

The posterior of \mathbf{x} transformed into itself, $p(\varphi|\mathbf{x}, \mathbf{x}, \mu, \kappa = \mathbf{0}) = \prod_j \mathcal{M}(\varphi_j|\hat{\mu}_j, \hat{\kappa}_j)$ gives a probabilistic description of the stabilizer of \mathbf{x} , and hence must be invariant. Clearly, the angle between \mathbf{x} and \mathbf{x} is zero, so $\hat{\mu} = \mathbf{0}$. On the other hand, $\hat{\kappa}$ contains information about \mathbf{x} and is invariant. To see this, recall that $\hat{\kappa}_j = \|\hat{\eta}_j\|$. Using eq. 12 we obtain $\hat{\kappa}_j = \|\mathbf{u}_j\|^2 \sigma^{-2} = \|\mathbf{W}_j^T \mathbf{x}\|^2 \sigma^{-2}$. Since every transformation \mathbf{Q} in the learned toroidal group acts on the 2D vector \mathbf{u}_j by rotation, the length of \mathbf{u}_j is left invariant.

We recognize the computation of $\hat{\kappa}$ as the *square pooling* operation often applied in convolutional networks to gain invariance: project an image onto filters \mathbf{W}_{2j-1} and \mathbf{W}_{2j} , and sum the squares. This computation follows as a necessary consequence of our model setup. In section 3.3, we will find that the model for non-maximal tori is even more informative about the proper pooling scheme.

Since we want to use $\hat{\kappa}$ as an invariant representation, we should try to find an appropriate metric on $\hat{\kappa}$ -space. Let $\hat{\kappa}(\mathbf{x})$ be defined by $p(\varphi|\mathbf{x}, \mathbf{x}, \kappa = \mathbf{0}) = \prod_j \mathcal{M}(\varphi|\hat{\mu}_j, \hat{\kappa}_j(\mathbf{x}))$. We suggest using the Hellinger distance:

$$\begin{aligned} H^2(\hat{\kappa}(\mathbf{x}), \hat{\kappa}(\mathbf{y})) &= \frac{1}{2} \sum_j \left(\sqrt{\hat{\kappa}_j(\mathbf{x})} - \sqrt{\hat{\kappa}_j(\mathbf{y})} \right)^2 \\ &= \frac{1}{2\sigma^2} \sum_j \|\mathbf{u}_j\|^2 + \|\mathbf{v}_j\|^2 - 2\|\mathbf{u}_j\| \|\mathbf{v}_j\| \\ &= \frac{1}{2\sigma^2} \sum_j \|\mathbf{v}_j - \mathbf{R}(\hat{\mu}_j) \mathbf{u}_j\|^2, \end{aligned}$$

which is equal to the *exact manifold distance* (eq. 13) up to a factor of $\frac{1}{2\sigma^2}$. The first step of this derivation uses eq. 12 under a uniform prior ($\eta_j = \mathbf{0}$), while the second step again makes use of the fact that $\hat{\mu}_j$ is the angle between \mathbf{u}_j and \mathbf{v}_j so that $\|\mathbf{u}_j\| \|\mathbf{v}_j\| = \mathbf{u}_j^T \mathbf{R}(\hat{\mu}_j) \mathbf{v}_j$.

3.2. Relation to the Discrete Fourier Transform

We show that the DFT is a special case of TSA. The DFT of a discrete 1D signal $\mathbf{x} = (x_1, \dots, x_D)^T$ is defined:

$$X_j = \sum_{n=0}^{D-1} x_n \rho^{-jn} \quad (14)$$

where $\rho = e^{2\pi i/D}$ is the D -th primitive root of unity. If we choose a basis of sinusoids for the filters in \mathbf{W} ,

$$\begin{aligned} \mathbf{W}_{(:, 2j-1)} &= \Re(\rho^{-j}, \dots, \rho^{-j(D-1)})^T \\ &= (\cos(2\pi j/D), \dots, \cos(2\pi j(D-1)/D))^T \\ \mathbf{W}_{(:, 2j)} &= \Im(\rho^{-j}, \dots, \rho^{-j(D-1)})^T \\ &= (\sin(-2\pi j/D), \dots, \sin(-2\pi j(D-1)/D))^T, \end{aligned} \quad (15)$$

then the change of basis performed by \mathbf{W} is a DFT. Specifically, $\Re(X_j) = \mathbf{u}_{j_1}$ and $\Im(X_j) = \mathbf{u}_{j_2}$.

Now suppose we are interested in the transformation taking some arbitrary fixed vector $\mathbf{e} = \mathbf{W}(1, 0, \dots, 1, 0)^T$ to \mathbf{x} . The posterior over φ_j is $p(\varphi_j|\mathbf{e}, \mathbf{x}, \eta_j = \mathbf{0}, \sigma = 1) = \mathcal{M}(\varphi_j|\hat{\eta}_j)$, where (by eq. 12) we have $\hat{\eta}_j = \|\mathbf{u}_j\| [\cos(\theta_j), \sin(\theta_j)]^T$, θ_j being the angle between \mathbf{u}_j and the ‘‘real axis’’ $\mathbf{e}_j = (1, 0)^T$. In conventional coordinates, the precision of the posterior is equal to the modulus of the DFT, $\hat{\kappa}_j = \|\mathbf{u}_j\| = |X_j|$, and the mean of the posterior is equal to the phase of the Fourier transform, $\hat{\mu} = \theta_j = \arg(X_j)$. Therefore, TSA provides a probabilistic interpretation of the DFT, and makes it possible to learn an appropriate generalized transform from data.

3.3. Modeling a Lie subalgebra

Typically, one is interested in learning groups with fewer than J degrees of freedom. Since the stabilizer representation is invariant to all transformations from the maximal torus which contains the symmetry group of interest, it is certainly also invariant to the subgroup of the maximal torus that is the true symmetry group. However, this representation will identify vectors that are in fact distinct from the point of view of the true symmetry group.

As we have seen, for one parameter compact subgroups of a maximal torus, the weights of the irreducible representations must be integers. We model this using a coupled rotation matrix, as follows:

$$\mathbf{Q}(s) = \mathbf{W} \begin{bmatrix} \mathbf{R}(\omega_1 s) & & \\ & \ddots & \\ & & \mathbf{R}(\omega_J s) \end{bmatrix} \mathbf{W}^T \quad (16)$$

Where $s \in [0, 2\pi]$ is the scalar parameter of this subgroup. The likelihood then becomes $\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mathbf{Q}(s)\mathbf{x}, \sigma^2)$.

The right prior for this likelihood is the *generalized von-Mises* (Gatto & Jammalamadaka, 2007):

$$p(s) = \mathcal{M}^+(s|\eta^+) = \exp(\eta^+ \cdot T^+(s)) / Z^+ \\ = \exp\left(\sum_{j=1}^K \kappa_j^+ \cos(js - \mu_j^+)\right) / Z^+$$

where $T^+(s) = [\cos(s), \sin(s), \dots, \cos(Ks), \sin(Ks)]^T$.

Using similar reasoning as in the non-coupled case, we find that this prior is conjugate to the coupled likelihood, so that $p(s|\mathbf{x}, \mathbf{y}) \propto \exp(\hat{\eta}^+ \cdot T^+(s))$, with:

$$\hat{\eta}_j^+ = \eta_j^+ + \sum_{k:\omega_k=j} \hat{\eta}_k \quad (17)$$

where $\hat{\eta}_k$ is obtained from eq. 12 using a uniform prior $\eta_k = \mathbf{0}$. The sum in this update equation performs a pooling operation over a *weight space*, meaning those invariant subspaces k whose weight $\omega_k = j$. In fact, the norm of any linear combination of same-weight representations is invariant, and the maximum of any two $\hat{\eta}_j$ or $\hat{\eta}_j^+$ is also invariant (Kanatani, 1990). The similarity to sum-pooling and max-pooling in convnets is striking, but whereas in convnets the filters adapt to a fixed pooling scheme, a TSA model can adapt the representation weights ω so as to perform the right kind of pooling.

In the non-coupled model, there are $J = D/2$ degrees of freedom in the group and the invariant representation is $D - J = J$ -dimensional ($\hat{\kappa}_1, \dots, \hat{\kappa}_J$). For the coupled model, there is only one degree of freedom in the group, so the invariant representation should be $D - 1$ dimensional. In the mathematically clean case where all ω_k are distinct, we have J variables $\kappa_1^+, \dots, \kappa_J^+$ that are invariant. Furthermore, from eq. 16 we see that as \mathbf{x} is transformed, the angle between \mathbf{x} and an arbitrary fixed reference vector in subspace j transforms as $\theta_j(s) = \delta_j + \omega_j s$ for some data-dependent initial phase δ_j . It follows that $\omega_j \theta_k(s) - \omega_k \theta_j(s) = \omega_j(\delta_k + \omega_k s) - \omega_k(\delta_j + \omega_j s) = \omega_j \delta_k - \omega_k \delta_j$ is invariant. In this way, we can easily construct another $J - 1$ invariants, but unfortunately these are not stable because the angle estimates can be inaccurate for low-energy subspaces. We leave the problem of devising a stable representation and a sensible metric on this space for future work.

The normalization constant Z^+ for the GvM has so far only been described for the case of $K = 2$ harmonics, but we have found a closed form solution in terms of the so-called modified Generalized Bessel Functions (GBF) of K -variables $\kappa^+ = \kappa_1^+, \dots, \kappa_K^+$ and parameters $\exp(-i\mu^+) = \exp(-i\mu_1^+), \dots, \exp(-i\mu_K^+)$ (Dattoli et al., 1991):

$$Z^+(\kappa^+, \mu^+) = 2\pi I_0(\kappa^+; e^{-i\mu^+}). \quad (18)$$

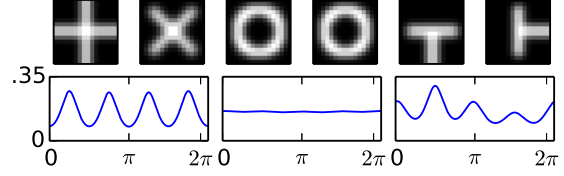


Figure 2. Posterior distribution over s for three image pairs.

We have developed a novel, highly scalable algorithm for the computation of GBF of many variables, which is described in the supplementary material.

Figure 2 shows the posterior over s for three image pairs related by different rotations and containing different symmetries. The weights \mathbf{W} and ω were learned by the procedure described in the next section. It is quite clear from this figure that MAP inference does not give a complete description of the possible transformations relating the images when the images have a degree of rotational symmetry. The posterior distribution of our model provides a sensible way to deal with this kind of uncertainty, which (in the case of 2D translations) is at the heart of the well known aperture problem in vision. Thus, our model could be used to estimate “multimodal optical flows”. Having a tractable posterior is also particularly important if the model is to be used to estimate longer sequences (akin to HMM/LDS models, but non-linear), where one may encounter multiple high-density trajectories.

If required, accurate MAP inference can be performed using the algorithm of Sohl-Dickstein et al. (2010), as described in the supplementary material. This allows us to compute the exact manifold distance for the coupled model.

3.4. Maximum Marginal Likelihood Learning

We train the model by gradient descent on the marginal likelihood. Perhaps surprisingly given the non-linearities in the model, the integrations required for the evaluation of the marginal likelihood can be obtained in closed form for both the coupled and decoupled models. For the decoupled model we obtain:

$$p(\mathbf{y}|\mathbf{x}) = \int_{\varphi \in \mathbb{T}^J} \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}) \prod_j \mathcal{M}(\varphi_j|\eta_j) d\varphi \\ = \frac{\exp(-\frac{1}{2\sigma^2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2))}{\sqrt{(2\pi\sigma)^D}} \prod_j \frac{I_0(\hat{\kappa}_j)}{I_0(\kappa_j)}. \quad (19)$$

Observing that $I_0(\hat{\kappa}_j)/I_0(\kappa_j)$ is the ratio of normalization constants of regular von-Mises distributions, the analogous expression for the coupled model is easily seen to be equal to eq. 19, only replacing $\prod_j I_0(\hat{\kappa}_j) / I_0(\kappa_j)$ by $Z^+(\hat{\kappa}^+, \hat{\mu}^+)/Z^+(\kappa^+, \mu^+)$. The derivation of this result

can be found in the supplementary material.

The gradient of the log marginal likelihood of the uncoupled model w.r.t. a batch \mathbf{X}, \mathbf{Y} (both storing N vectors in the columns) is:

$$\begin{aligned} \frac{d}{d\mathbf{W}} \ln p(\mathbf{Y}|\mathbf{X}) = & \mathbf{X}(\mathbf{R}^T(\mu)\mathbf{W}^T\mathbf{Y} \succ \mathbf{A} + \mathbf{W}^T\mathbf{X} \succ \mathbf{B}^y)^T \\ & + \mathbf{Y}(\mathbf{R}(\mu)\mathbf{W}^T\mathbf{X} \succ \mathbf{A} + \mathbf{W}^T\mathbf{Y} \succ \mathbf{B}^x)^T. \end{aligned}$$

where we have used $(\mathbf{P} \succ \mathbf{Q})_{2j,n} = \mathbf{Q}_{j,n}\mathbf{P}_{2j,n}$ and $(\mathbf{P} \succ \mathbf{Q})_{2j-1,n} = \mathbf{Q}_{j,n}\mathbf{P}_{2j-1,n}$ as a ‘‘subspace weighting’’ operation. $\mathbf{A}, \mathbf{B}^{(x)}$ and $\mathbf{B}^{(y)}$ are $D \times N$ matrices with elements

$$\begin{aligned} a_{jn} &= \frac{I_1(\hat{\kappa}_{jn})\kappa_j}{I_0(\hat{\kappa}_{jn})\hat{\kappa}_{jn}\sigma^2}, \\ b_{jn} &= \frac{\|\mathbf{W}_j\mathbf{y}^{(n)}\|^2}{\kappa_j\sigma^2}, \end{aligned}$$

where the $\hat{\kappa}_{jn}$ is the posterior precision in subspace j for image pair $\mathbf{x}^{(n)}, \mathbf{y}^{(n)}$ (the n -th column of \mathbf{X} , resp. \mathbf{Y}).

The gradient of the coupled model is easily computed using the differential recurrence relations that hold for the GBF (Dattoli et al., 1991).

We use minibatch Stochastic Gradient Descent (SGD) on the log-likelihood of the uncoupled model. After every parameter update, we orthogonalize \mathbf{W} by setting all singular values to 1: Let $\mathbf{U}, \mathbf{S}, \mathbf{V} = \text{svd}(\mathbf{W})$, then set $\mathbf{W} := \mathbf{UV}$. This procedure and all previous derivations still work when the basis is undercomplete, i.e. has fewer columns (filters) than rows (dimensions in data space). To learn ω_j , we estimate the relative angular velocity $\omega_j = \theta_j/\delta$ from a batch of image patches rotated by a sub-pixel amount $\delta = 0.1^\circ$.

4. Experiments

We trained a TSA model with 100 filters on a stream of 250,000 16×16 image patches $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}$. The patches $\mathbf{x}^{(t)}$ were drawn from a standard normal distribution, and $\mathbf{y}^{(t)}$ was obtained by rotating $\mathbf{x}^{(t)}$ by an angle s drawn uniformly at random from $[0, 2\pi]$. The learning rate α was initialized at $\alpha_0 = 0.25$ and decayed as $\alpha = \alpha_0/\sqrt{T}$, where T was incremented by one with each pass through the data. Each minibatch consisted of 100 data pairs. After learning \mathbf{W} , we estimate the weights ω_j and sort the filter pairs by increasing absolute value for visualization. As can be seen in fig. 3, the filters are very clean and the weights are estimated correctly except for a few filters on row 1 and 2 that are assigned weight 0 when in fact they have a higher frequency.

We tested the utility of the model for invariant classification on a rotated version of the MNIST dataset, using a 1-Nearest Neighbor classifier. Each digit was rotated by a random angle and rescaled to 16×16 pixels, resulting in

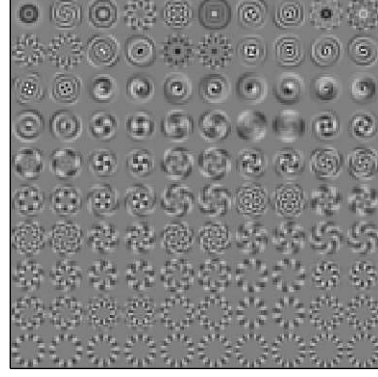


Figure 3. Filters learned by TSA, sorted by absolute frequency $|\omega_j|$. The learned ω_j -values range from -11 to 12 .

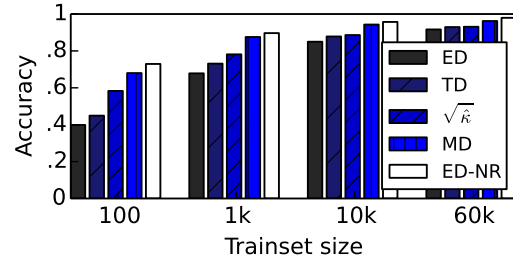


Figure 4. Results of classification experiment. See text for details.

60k training examples and 10k testing examples, with no rotated duplicates. We compared the Euclidean distance (ED) in pixel space, tangent distance (TD) (Simard et al., 2000), Euclidean distance on the space of $\sqrt{\hat{\kappa}}$ (equivalent to the exact manifold distance for the maximal torus, see section 3.1), the true manifold distance for the 1-parameter 2D rotation group (MD), and the Euclidean distance on the non-rotated version of the MNIST dataset (ED-NR). The results in fig. 4 show that TD outperforms ED, but is outperformed by $\sqrt{\hat{\kappa}}$ and MD by a large margin. In fact, the MD-classifier is about as accurate as ED on a much simpler dataset, demonstrating that it has almost completely modded out the variation caused by rotation.

5. Conclusions and outlook

We have presented a novel principle for learning disentangled representations, and worked out its consequences for a simple type of symmetry group. This leads to a completely tractable and very elegant model with potential applications to invariant classification and Bayesian estimation of motion. The model reproduces the pooling operations of convolutional networks from a probabilistic and Lie-group theoretic perspective, and provides a probabilistic interpretation of the DFT and its generalizations.

The type of disentangling obtained in this paper is contingent upon the rather minimalist assumption that all that can be said about images is that they are equivalent (rotated copies) or inequivalent. However, the universal nature of Weyl's principle bodes well for future applications in deep, non-linear and non-commutative disentangling.

References

- Bengio, Yoshua and Lecun, Yann. International Conference on Learning Representations, 2014. URL <https://sites.google.com/site/representationlearning2014/>.
- Bengio, Yoshua, Courville, Aaron, and Vincent, Pascal. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*, pp. 1–30, February 2013. ISSN 1939-3539.
- Bethge, Matthias, Gerwinn, Sebastian, and Macke, Jakob H. Unsupervised learning of a steerable basis for invariant image representations. *Proceedings of SPIE Human Vision and Electronic Imaging XII (E1105)*, February 2007.
- Cadieu, Charles F and Olshausen, Bruno A. Learning intermediate-level representations of form and motion from natural movies. *Neural computation*, 24(4):827–66, April 2012. ISSN 1530-888X.
- Dattoli, G., Chiccoli, C., Lorenzutta, S., Maino, G., Richetta, M., and Torre, A. A Note on the Theory of n-Variable Generalized Bessel Functions. *Il Nuovo Cimento B*, 106(10):1159–1166, October 1991. ISSN 0369-3554.
- Gatto, Riccardo and Jammalamadaka, SR. The generalized von Mises distribution. *Statistical Methodology*, 4(3): 341–353, 2007.
- Goodfellow, I, Le, Q, and Saxe, A. Measuring invariances in deep networks. *Advances in Neural Information Processing Systems*, pp. 1–9, 2009.
- Kanatani, Kenichi. *Group Theoretical Methods in Image Understanding*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1990. ISBN 0387512535.
- LeCun, Y and Bottou, L. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
- Mardia, Kanti V. and Jupp, Peter E. *Directional Statistics*. John Wiley & Sons, 1 edition, 1999.
- Memisevic, Roland. On multi-view feature learning. *International Conference on Machine Learning*, 2012.
- Memisevic, Roland and Hinton, Geoffrey E. Learning to Represent Spatial Transformations with Factored Higher-Order Boltzmann Machines. *Neural Computation*, 22(6):1473–1492, June 2010. ISSN 1530-888X. doi: 10.1162/neco.2010.01-09-953.
- Miao, Xu and Rao, Rajesh P N. Learning the Lie groups of visual invariance. *Neural computation*, 19(10):2665–93, October 2007. ISSN 0899-7667. doi: 10.1162/neco.2007.19.10.2665.
- Rao, RPN and Ruderman, DL. Learning Lie groups for invariant visual perception. *Advances in neural information processing systems*, 816:810–816, 1999.
- Simard, Patrice Y., Le Cun, Yann a., Denker, John S., and Victorri, Bernard. Transformation invariance in pattern recognition: Tangent distance and propagation. *International Journal of Imaging Systems and Technology*, 11(3):181–197, 2000. ISSN 08999457. doi: 10.1002/1098-1098(2000)11:3<181::AID-IMA1003>3.3.CO;2-5.
- Soatto, Stefano. Actionable information in vision. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2138–2145. IEEE, September 2009. ISBN 978-1-4244-4420-5.
- Sohl-Dickstein, J, Wang, JC, and Olshausen, BA. An unsupervised algorithm for learning lie group transformations. *arXiv preprint*, 2010.
- Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008. doi: 10.1145/1390156.1390294.
- Welling, Max, Rosen-zvi, Michal, and Hinton, Geoffrey. Exponential Family Harmoniums with an Application to Information Retrieval. *Advances in neural information processing systems*, 17, 2005.